Submitted to the Journal of Biomedical Informatics (Special issue on Auditing Terminologies).
draft material -- please do not cite

A Graph-based Approach to Auditing RxNorm

Olivier Bodenreider[1$], *Lee B. Peters*[1]

[1] Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD

[$]Corresponding author:

Dr. Olivier Bodenreider

National Library of Medicine

8600 Rockville Pike - MS 3841 (Bldg 38A, Rm B1N28U)

Bethesda, MD 20894 - USA

phone: 301 435-3246 - fax: 301 480-3035

olivier@nlm.nih.gov

Abstract

Objectives: RxNorm is a standardized nomenclature for clinical drugs developed by the National Library of Medicine. In this paper, we audit relations in RxNorm for consistency and completeness through the systematic analysis of the graph of its concepts and relationships.

Methods: The representation of generic drugs is normalized in order to make it compatible with that of branded drugs. All meaningful paths between two nodes in the type graph are computed and instantiated. Alternate paths are automatically compared and manually inspected in case of inconsistency.

Results: The 115 meaningful paths identified in the graph type can be grouped into 28 groups with respect to start and end nodes. Of the 19 groups of alternate paths (i.e., with two or more paths) between the start and end nodes, 8 (42%) exhibit inconsistencies. Overall, 15 (13%) of the 115 paths are inconsistent with other alternate paths. 240 errors were identified and reported to the RxNorm team.

Conclusion: The errors identified involve missing nodes (93), missing links (17), extraneous links (129) and one case of mix-up between two ingredients. Our auditing method proved both sensitive and specific. Some recommendations for the development of RxNorm are provided.

Keywords

Biomedical terminologies, Auditing methods, RxNorm, graphs.

1 Introduction

Terminology development in biomedicine largely relies on the manual work of human editors (sometimes called modelers) [e.g., 1, 2]. Although sometimes facilitated by the use of knowledge representation formalisms such as description logics, this process is known to be error-prone [3, 4]. Many approaches have been proposed for analyzing large biomedical terminologies, based on the property of their terms [3-6], on their structure [7-10] and on their semantics [3, 4, 11, 12]. Most approaches focus on auditing hierarchical relations, which form the backbone of biomedical terminologies [7, 8, 10]. Many terminology developers include quality assurance and quality control processes as part of the development cycle [13]. However, such mechanisms fail to capture many errors and independent researchers and the user community play an important role in identifying and reporting errors in biomedical terminologies.

From a structural perspective, most biomedical terminologies can be seen as directed graphs in which nodes are concepts and links are semantic relationships. A path between two concepts can be characterized by the sequence of relationships that need to be traversed in order to reach a target concept from a source concept. While broad terminologies (e.g., SNOMED CT, NCI Thesaurus) usually have a complex model of meaning (or T-Box in description logics-based terminologies), specialized terminologies such as RxNorm (presented in detail later) only define few major categories (or types) in the domain and their interrelations. In such terminologies, most of the assertions hold among instances of these categories (rather than among the categories themselves) and are associative rather than hierarchical. The graph of types provides a model against which graphs of instances can be validated. For example, all paths defined in the model are expected to be instantiated, allowing to check for completeness (of nodes and links at the instance level). Similarly, alternate paths between two types known to be consistent at the type level are also expected to be consistent at the instance level.

The objective of this study is to audit relations in RxNorm for consistency and completeness through the systematic analysis of the graph of its concepts and relationships (at the instance level, in reference to the type level). More specifically, we hypothesize that the traversal of equivalent paths yielding different results is indicative of errors in the graph of instances, including missing links and erroneous links, and, possibly, missing nodes.

2 Background: RxNorm

RxNorm is a standardized nomenclature for clinical drugs developed by the National Library of Medicine [14]. RxNorm is one of a suite of designated standards for use in U.S. Federal Government systems for the electronic exchange of clinical health information. RxNorm has been used as part of a mediation strategy to exchange medication data between the Veterans Affairs (VA) and the Department of Defense (DoD) clinical information systems [15] and as a drug vocabulary for personal health records [16]. It is also expected to become an enabling resource for applications such as e-prescribing [17] and medication reconciliation [18].

2.1 RxNorm categories

The RxNorm data set is organized around eight major categories, called "term types" in RxNorm parlance. There are four categories for generic drugs and four equivalent categories for branded drugs. The four categories for generic drugs are for ingredient alone (ingredient), ingredient plus strength (clinical drug component), ingredient plus dose form (clinical drug form) and ingredient plus strength and dose form (clinical drug). Analogously, the four categories for brand name drug concepts (referred to hereafter as branded concepts) are brand name (alone), branded drug component (brand name plus strength), branded drug form (brand name plus dose form) and branded drug (brand name plus strength and dose form). Table 1 lists the eight major categories and some instances. The dataset under investigation in this study (April 1, 2008) comprises, after excluding obsolete data, 3,460 ingredients (ignoring specific salts), 9,740 brand names, 13,362 clinical drug components, 13,868 branded drug components, 18,097 clinical drugs, 14,539 branded drugs, 8,160 clinical drug forms and 11,376 branded drug forms.

2.2 RxNorm relations

As shown in Figure 1, relations are defined among branded concepts and among generic concepts. For each brand name concept, there exist one or more branded drug components, branded drug names and branded drug forms. Each ingredient is associated with one or more clinical drug components, clinical drugs and clinical drug forms. Moreover, the RxNorm drug entities are related to each other by a well-defined set of named relationships. For example, brand name concepts are related to branded drug component concepts by the relationships *ingredient_of* and *has_ingredient*, the latter being the inverse relationship. Examples of relationships at the instance level include:

- *Zyrtec 5 MG Oral Tablet consist_of Cetirizine 5 MG [Zyrtec]*
- *Cetirizine 5 MG constitutes Cetirizine 5 MG Oral Tablet*

Figure 1 shows all relationships between the various kinds of drug entities. It must be noted that the relationship *isa* defined between branded drug and branded drug component and between *clinical drug* and *clinical drug component* does not have the usual semantics of the subsumption relation of the same name, but simply links an entity with ingredient (resp. brand name), strength and dose form to the corresponding entity with ingredient (resp. brand name) and dose form, but no strength.

In addition to relations among branded concepts and among generic concepts, RxNorm also defines relations between branded concepts and generic concepts. As illustrated in Figure 1, most relations are between entities at the same level (e.g., ingredient plus strength to brand name plus strength). This relationship is called *tradename_of* from branded concepts to generic concepts, the inverse relationship being *has_tradename*. Additionally, RxNorm defines the relationship *consists_of* between branded drugs and clinical drug components, with *constitutes* as its inverse. Examples of relationships at the instance level include:

- *Cetirizine has_tradename Zyrtec*
- *Cetirizine 10 MG Oral Tablet [Zyrtec] tradename_of Cetirizine 10 MG Oral Tablet*

While all branded concepts stand in a relation to some generic drug concepts, some generic drug concepts are not linked to any branded concepts. For example, there is no branded concept corresponding to *Cetirizine 10 MG Extended Release Tablet*, which means that this particular ingredient, strength and dose form combination is not commercialized under a particular brand, but rather available as a generic drug.

All relations in RxNorm are represented bidirectionally, i.e., for each relation (e.g., *Cetirizine has_tradename Zyrtec*), the inverse relation (i.e., *Zyrtec tradename_of Cetirizine*) is also recorded in the RxNorm dataset.

One major difference between the representation of generic and branded concepts is that each branded drug, branded drug component and branded drug form is related to only one brand name, whereas, for their generic counterpart, multi-ingredient clinical drugs and clinical drug forms are related to multiple ingredients and clinical drug components. As shown in Figure 2, the branded drug (*Sulfamethoxazole 400 MG / Trimethoprim 80 MG Oral Tablet [Bactrim]*) is linked to one clinical drug (*Sulfamethoxazole 400 MG / Trimethoprim 80 MG Oral Tablet*). However, the branded drug is linked to one branded drug component (*Sulfamethoxazole 400 MG / Trimethoprim 80 MG [Bactrim]*), whereas the clinical drug is linked to two clinical drug components (*Sulfamethoxazole 400 MG* and *Trimethoprim 80 MG*), one for each ingredient of the multi-ingredient drug (*Sulfamethoxazole* and *Trimethoprim*).

2.3 RxNorm Web Services API

A browser called RxNav[1] was developed in 2004 to access the RxNorm data set and display graphically all related concepts and the relations between them [16]. RxNav uses web services to access the RxNorm data. In early 2008, the web services that access the RxNorm data were enhanced and made available publicly [19]. The current application programming interface (API) comprises functions for resolving drug names and codes into RxNorm identifiers, for accessing the properties of drug concepts, and for getting the related concepts of RxNorm entities. Here, we take advantage of the latter set of functions for exploring the RxNorm graph computationally.

3 Methods and Results

The methods used in this study can be summarized as follows. We start by creating a normalized representation of generic drug entities mirroring the representation on the brand side. Then, we identify all meaningful paths between two categories, for all the instances of the source category. Finally, we assess the consistency of alternate paths between pairs of categories by comparing sets of instances reached through the various alternate paths.

3.1 Normalizing the representation of generics drug entities with respect to branded drug entities

As explained earlier, the representation of multi-ingredient drugs differs in RxNorm for generic drug entities compared to branded drug entities. While useful for common uses of RxNorm as there is no such thing in practice as a combination of ingredients, we found this difference to be a hindrance to our

[1] http://mor.nlm.nih.gov/download/rxnav/

auditing endeavor. We hypothesized that our exploration of the RxNorm graph would be simplified if the same algorithms could be used for exploring both single- and multi-ingredient drugs.

As illustrated by the differences between Figure 2 and Figure 3, normalization occurs at the level of ingredients and clinical drug components and their relations to other generic drug entities, namely clinical drugs (for clinical drug components) and ingredients (for clinical drug forms), as well as to the corresponding branded drug entities, namely brand names (for ingredients) and branded drug components (for clinical drug components). The normalization process simply reifies multi-ingredient entities into single-ingredient entities.

In practice, the normalization process creates new ingredient concepts for combinations of ingredients and new clinical drug component concepts for combinations of clinical drug components. For example, as shown in Figure 3, the two ingredients of the brand name *Bactrim*, *Sulfamethoxazole* and *Trimethoprim*, are grouped into the new ingredient concept *Sulfamethoxazole / Trimethoprim*. Similarly, the two clinical drug components of the branded drug component *Sulfamethoxazole 400 MG / Trimethoprim 80 MG [Bactrim]*, *Sulfamethoxazole 400 MG* and *Trimethoprim 80 MG*, are grouped into the new clinical drug component concept *Sulfamethoxazole 400 MG / Trimethoprim 80 MG*. The relations of the newly created concepts are adapted accordingly. A single link is created from the new ingredient *Sulfamethoxazole / Trimethoprim* to both the clinical drug form *Sulfamethoxazole / Trimethoprim Oral Tablet* and the brand name *Bactrim*. Similarly, a single link is created from the new clinical drug component *Sulfamethoxazole 400 MG / Trimethoprim 80 MG* to both the clinical drug *Sulfamethoxazole 400 MG / Trimethoprim 80 MG Oral Tablet* (*ingredient_of*) and the branded drug component *Sulfamethoxazole 400 MG / Trimethoprim 80 MG [Bactrim]* (*tradename _of*). Finally, a single link is also created between the new ingredient and the new clinical drug component (*ingredient_of*). All links are represented bidirectionally. The original links are removed, and so are the original ingredients and clinical drug components if they do not participate in any other single- or multi-ingredient drug entities.

3.2 Identifying all meaningful paths

Starting from any of the eight major categories in RxNorm, we want to explore all paths to any of the seven other categories. In order to avoid identifying meaningless paths, we constrain the traversal of the graph in the following ways:

- Each node is allowed to be traversed only once

  The rationale is simply to avoid circuits.

- The path shall only cross once between the generic and branded drug entities

  One reason for this is that *some generic drug entities do not have any associated branded drug entities*. Therefore, it would be predictably inconsistent to go, from example, from one ingredient to the corresponding clinical drugs through branded drug entities. There is no branded drug entity for the clinical drug *Cetirizine 10 MG Extended Release Tablet*, discussed earlier. Therefore, paths from *Cetirizine* to clinical drugs on the generic side will find *Cetirizine 10 MG Extended Release Tablet*, while paths going through the brand side will not.

The other reason is that *several brands may produce the same drugs containing the same ingredient(s)*. Therefore, it would also be predictably inconsistent to go, from example, from one clinical drug component (specific to one particular brand) to the corresponding brand name(s) through generic drug entities, as the specificity of the original brand would be lost, by design, on the generic side.

Additionally, the graph traversal was further constrained in order to avoid the exploration of predictably inconsistent paths. These constraints include:

- Entities such as ingredient and brand name shall not be traversed from and to any other entities bearing strength or dose form properties

  The entities ingredient and brand name do not contain any strength or dose form information. Therefore, going through such an entity from an entity that contains strength information to an entity that contains dose form information results in wrongly associating every strength of the source entity with every dose form of the target entity. For example, doing so would result in the wrong association of *Cetirizine 10 MG* with *Cetirizine Oral Solution* (10 MG is not an appropriate strength for oral solutions).

- For paths starting on the brand side and crossing over to the generic side, any property (strength or dose form) of the target entity, if acquired along the path, shall be acquired on the brand side

  As mentioned earlier, *some generic drug entities do not have any associated branded drug entities*. Therefore, acquiring the property on the generic side might result in reaching generic drug entities that do not correspond to the source (brand) entity. For example, when going from the branded drug component *Cetirizine 10 MG [Zyrtec]* to clinical drugs through the clinical drug component *Cetirizine 10 MG* (i.e., acquiring the dose form property on the generic side), the clinical drug *Cetirizine 10 MG Extended Release Tablet* will be found, although there is no branded equivalent for this clinical drug. In contrast, going through the branded drugs *Zyrtec 10 MG Chewable Tablet* and *Zyrtec 10 MG Oral Tablet* (i.e., acquiring the dose form property on the brand side) will lead to the clinical drugs *Cetirizine 10 MG Chewable Tablet* and *Cetirizine 10 MG Oral Tablet*, which is correct.

- For paths starting on the generic side and crossing over to the brand side, any property (strength or dose form) of the source entity, if removed along the path, shall be removed on the brand side

  Not all brands produce all strengths and dose forms of a given drug. Therefore, removing a property on the generic side might result in reaching branded drug entities that do not correspond to the specific strength or dose form of the source (generic) entity. For example, as shown in Figure 4, when going from the clinical drug *Warfarin 1 MG Oral Tablet* to clinical drugs through the clinical drug form *Warfarin Oral Tablet* (i.e., removing the strength property on the generic side), four branded drug forms will be found (*Warfarin Oral Tablet [Coumadin]*, *Warfarin Oral Tablet [Jantoven]*, *Warfarin Oral Tablet [Marfarin]* and *Warfarin Oral Tablet [Warfin]*), although there are only two clinical drugs with a strength of 1MG for this form (*Coumadin 1 MG Oral Tablet* and *Jantoven 1 MG Oral Tablet*). In contrast, going through the branded drugs *Coumadin 1 MG Oral*

Tablet and *Jantoven 1 MG Oral Tablet* (i.e., removing the strength property on the brand side) will lead to the clinical drug components *Warfarin Oral Tablet [Coumadin]* and *Warfarin Oral Tablet [Jantoven]*, which is correct.

These constraints were easily implemented through a regular expressions applied on the sequence of transitions for a given path and to the sequence of states (i.e., lists of properties) for all the nodes in a path, emulating a finite state automaton.

Out of the 790 paths obtained after implementing the first two constraints, a total of 230 meaningful paths remain after the last three constraints have been applied. Since all relations in RxNorm are recorded bidirectionally, we actually have 115 pairs of inverse paths. Only one copy needs to be explored for each path pair. As shown in Table 2, these 115 paths can be grouped into 28 classes with respect to source and target nodes in the path.

3.3 Exploring meaningful paths

Each of the 115 meaningful paths (of categories) was explored as follows. Starting from the category corresponding to the first node in the path (source category), all instances of this node were retrieved. For each instance of the source category, we recorded how many instances of the target category could be reached, following the links indicated in the path of categories. The total number of instances reached for a given path is the sum of the number of target instances reached from each source instance. Equivalent paths are deemed consistent if the same set of target instances is reached from all alternate paths from the set of equivalent paths.

For example, there are three meaningful paths between clinical drug component (SCDC) and branded drug component (SBDC):

1. SCDC=>SBDC
2. SCDC=>SBD[2]=>SBDC
3. SCDC=>SCD=>SBD=>SBDC

Exploring path 3, the list of instances of SCDC (source instances) includes *Warfarin 1 MG*. As shown in Figure 5, the only SCD instance that can be reached from *Warfarin 1 MG* through the relationship *constitutes* is *Warfarin 1 MG Oral Tablet*. From this SCD instance, following the relationship *has_tradename*, the following SBD instances can be reached: *Coumadin 1 MG Oral Tablet* and *Jantoven 1 MG Oral Tablet*. The *Coumadin 1 MG Oral Tablet* leads to the SBDC *Warfarin 1 MG [Coumadin]* (target category) through the relationship *consists_of*. Similarly, *Jantoven 1 MG Oral Tablet* leads to the SCDC instance *Warfarin 1 MG [Jantoven]*. In summary, the source SCDC instance *Warfarin 1 MG* leads to two target SBDC instance *Warfarin 1 MG [Coumadin]* and *Warfarin 1 MG [Jantoven]* through the path SCDC=>SCD=>SBD=>SBDC. The source SCDC instance *Warfarin 1 MG* therefore contributes 2 target SBDC instances to the path SCDC=>SCD=>SBD=>SBDC. Overall, this path yields 13, 868 target

[2] The reader is referred to Table 1for the abbreviation of RxNorm categories used in path specifications

instances. Paths 1 and 2 above also yield the same 13,868 target instances and therefore the 3 alternate paths between SCDC and SBDC are deemed equivalent.

The RxNorm API was used to explore the paths. In particular, the function *getRelatedByRelationship()* was used for querying the instances of a given type that could be reached from a given RxNorm entity (instance) through a given link.

The results of the exploitation of the 115 meaningful paths are summarized in Table 2. In order to reduce the amount of information in this table, we only display one path for each set of equivalent paths. For example, from the three equivalent paths presented above for the start node SCDC and end node SBDC, column 3 confirms that there are indeed three paths, although only one of them (SCDC=>SBDC) is actually listed in column 4. In fact, column 5 indicates that there are two other unlisted equivalent paths for this path. Finally, column 6 lists the number of target instances reached for each set of equivalent paths.

Of the 28 groups of alternate paths expected to be equivalent, 9 groups contain only one path. Of the 19 groups having more than one path, all alternate paths are equivalent in 11 groups (58%), while 8 groups (42%) exhibit inconsistencies among alternate paths. Overall, 15 (13%) of the 115 paths are inconsistent with other alternate paths.

4 Discussion

4.1 Summary of inconsistencies

The analysis of Table 2 reveals that inconsistencies in three paths (BN=>SBD=>SBDF, IN=>BN and IN=>SCDC=>SCD=>SCDF) are actually responsible for the inconsistencies observed in the 15 paths. The reason for this is that the three paths are included as proper subpaths in the other 12 paths. For example, BN=>SBD=SBDF is a proper subpath of IN=>BN=>SBD=>SBDF from the group IN-SBDF.

The degree of inconsistency observed among alternate paths was usually small. For example, for the path IN-SCDF, the correct path yields 8,160 target instances, while the inconsistent alternate path yields 8,104 target instances.

4.2 Analysis of inconsistencies

Through manual analysis of the inconsistencies observed among alternate paths, this study revealed three major types of issues at the origin of the inconsistencies.

Type 1. These inconsistencies involved drug form entities (clinical or branded) linked to some ingredient or brand name, but not linked to a drug entity (clinical or branded).

- **Type 1a**. We found 36 cases of brand drug form concepts having no relation to any branded drug entity, but linked to some branded name concept. In this case, the direct path BN=>SBDF is inconsistent with the alternate path BN=>SBD=>SBDF. For example, the brand name *Sinemet* is related to the branded drug form *Carbidopa / Levodopa Oral Tablet [Sinemet]*, but neither concept

is linked to any branded drug entity. Errors of this type propagate directly to four other paths of which they are a proper subpath (e.g., IN=>BN=>SBD=>SBDF). Additionally, the same errors propagate indirectly to two other paths through the insertion of the direct relation SBD=>SCD (e.g., IN=>SN=>SBD=>SCD=>SBDF). Overall, this type of errors affects seven of the 15 alternate paths exhibiting inconsistencies.

- **Type 1b**. We found 57 cases of clinical drug form concepts having no relation to any clinical drug entity, but linked to some ingredient concept. In this case, the direct path IN=>SCDF is inconsistent with the alternate path IN=>SCDC=>SCD=>SCDF. For example, the clinical drug form *Papain Chewable Tablet* is related to the ingredient *Papain*, but is not linked to any clinical drug entity. Errors of this type propagate directly to five other paths of which they are a proper subpath (e.g., IN=>SCDC=>SCD=>SCDF=>SBDF=>SBD). Overall, this type of errors affects six of the 15 alternate paths exhibiting inconsistencies.

Type 2. The inconsistencies observed among paths from ingredient to brand name entities affect five of the 15 alternate paths exhibiting inconsistencies. Analyzing these inconsistencies revealed three distinct kinds of issues. In all three cases, the direct path BN=>IN is inconsistent with alternate paths such as BN=>SBDC=>SCDC=>IN.

- **Type 2a**. In 17 cases, a brand name entity has no relation to any ingredient entities. Examples include *Sochlor*, not linked directly to its ingredient, *Sodium chloride*.

- **Type 2b**. In 129 cases, a brand name entity has extraneous relations to some ingredient entities. Examples include *Benadryl Allergy Sinus*, inappropriately linked to the ingredient *Acetaminophen*, when its actual ingredients are *Diphenhydramine* and *Pseudoephedrine*.

- **Type 2c**. In 108 cases, a brand name entity refers to several combinations of ingredients depending on the dose form or strength of the drug. In such cases, a brand name entity is linked to multiple clinical drug components, each of which is linked to different sets of ingredients. For example, the brand name *Relagard* is linked to two branded drug component entities: *Acetic Acid 0.0092 MG/MG [Relagard]* and *Acetic Acid 0.009 MG/MG / Oxyquinoline Sulfate 0.00025 MG/MG [Relagard]*. Although sharing the same brand name, the former has only one ingredient (*Acetic Acid*), while the latter as 2 (*Acetic Acid* and *Oxyquinoline Sulfate*).

Type 3. What looks like a mix-up between two ingredients causes one inconsistency that is reflected in four paths. In this case, although the alternate paths exhibit the same numbers of target instances, the sets of target instances are actually different. The two ingredients involved in the mix-up are *Omega-3 Acid Ethyl Esters (USP)* and *Fatty Acids, Omega-3*. As nothing general is to be learned from this error, we do not report it here in detail.

Overall, the major types of errors identified in the RxNorm dataset include missing nodes (type 1a and 1b errors), missing relations (type 2a errors) and extraneous relations (type 2b errors).

4.3 Significance

The number of inconsistencies identified among alternate paths and the number of errors identified through their analysis is relatively modest (240, excluding inconsistencies of type 2c), which is a testimony to the high quality and careful curation of the RxNorm database. However, we believe this study is significant, because these errors are difficult to identify. In fact, these errors obviously defeated the quality assurance mechanisms currently in place in the RxNorm production system and had not been reported to (and acted upon by) the RxNorm team in the several years RxNorm has been available. We believe that only a systematic, principled analysis can identify such errors in a large dataset. The list of errors we identified was shared with the RxNorm developers.

Unlike other auditing methods, the graph-based process we developed for analyzing RxNorm groups errors in categories according to their origin. As a result, the errors reported to the RxNorm can be processed in groups and the appropriate quality assurance can be added to the production system.

Although the errors have not been reviewed by the RxNorm team yet, we believe our auditing method is both sensitive and specific. Sensitivity is illustrated by the errors of type 3, generated by one single mix-up between two ingredients. Inconsistencies of type 2c are probably not to be considered as errors, but rather as the consequences of an infelicitous representation mechanism for brand names associated with several combinations of ingredients (simply reflecting naming practices in the pharmaceutical industry). Otherwise, the other cases we reviewed correspond, in our opinion, to errors in the dataset.

4.4 Recommendations for the RxNorm development process

The normalization process developed for this study, which makes the representation of generic drug entities compatible with that of branded drug entities, is a critical element of our method. However, we do not recommend that the RxNorm developers change the current representation. In fact, reified combinations of ingredients and clinical drug component entities are artificial constructs, with no equivalent in the real world, and would therefore not be useful to most users of RxNorm.

Some of the errors detected in this study call for additional quality assurance processes to be implemented in the RxNorm production system. For example, it would be easy to check if a given clinical drug from with links to an ingredient is also linked to at least one clinical drug entity.

This study forced us to formalize what constitutes a meaningful path for traversing the RxNorm graph. Although a small number of constraints are required for ensuring meaningful traversal of the graph, we found it difficult to formulate these constraints. As the use of RxNorm increases, we suggest that guidance be added to the RxNorm documentation regarding traversal of the RxNorm graph.

Finally, the RxNorm graph contains some redundancy, but redundancy is not present systematically throughout the graph. On the one hand, it might be better to provide users with the minimal number of relations necessary for traversing the graph in a meaningful way. (This option would call for removing the direct relation between ingredient and clinical drug form, for example, as it can be reconstructed through the path IN=>SCDC=>SCD=>SCDF). On the other hand, it might be useful to some users to have a fully saturated set of relations. (This option would call for adding a direct relation between ingredient and clinical drug, mirroring the relation between brand name and branded drug on the brand side).

5 Conclusions

Through the graph-based method de developed for auditing RxNorm, we identified 240 errors, including missing nodes (93), missing links (17), extraneous links (129) and one case of mix-up between two ingredients. Our auditing method proved both sensitive and specific. Based on this analysis, we recommended some changes to the RxNorm quality assurance process, as well as additions to the RxNorm documentation.

Other approaches could be used to address the same issue, including role composition in a description logics-based environment. However, we found RxNorm to be amenable to graph-based approaches. Moreover, the availability of the RxNorm API allowed us to reduce low-level programming to a minimum.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

[1] Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. J Biomed Inform 2005;38(2):114-29.

[2] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662-6.

[3] Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods Inf Med 2005;44(4):498-507.

[4] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Medinfo 2004;11(Pt 1):482-6.

[5] Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L. The compositional structure of Gene Ontology terms. Pac Symp Biocomput 2004:214-25.

[6] Ogren PV, Cohen KB, Hunter L. Implications of compositionality in the gene ontology for its curation and usage. Pac Symp Biocomput 2005:174-85.

[7] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. J Biomed Inform 2003;36(6):450-61.

[8] Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artif Intell Med 2004;31(1):29-44.

[9] Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. J Biomed Inform 2007;40(5):561-81.

[10] Bodenreider O. Strength in numbers: exploring redundancy in hierarchical relations across biomedical terminologies. AMIA Annu Symp Proc 2003:101-5.

[11] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51.

[12] Cornet R, Abu-Hanna A. Two DL-based methods for auditing medical terminological systems. AMIA Annu Symp Proc 2005:166-70.

[13] Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. J Am Med Inform Assoc 2006;13(6):676-90.

[14] Liu S, Wei M, Moore R, Ganesan VAGV, Nelson SANS. RxNorm: prescription for electronic drug information exchange. IT Professional 2005;7(5):17-23.

[15] Bouhaddou O, Warnekar P, Parrish F, Do N, Mandel J, Kilbourne J, et al. Exchange of Computable Patient Data Between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): Terminology Standards Strategy. J Am Med Inform Assoc 2007.
[16] Zeng K, Bodenreider O, Kilbourne JT, Nelson SJ. RxNav: Towards an integrated view on drug information. Medinfo 2007:P386.
[17] Schade CP, Sullivan FM, de Lusignan S, Madeley J. e-Prescribing, efficiency, quality: lessons from the computerization of UK family practice. J Am Med Inform Assoc 2006;13(5):470-5.
[18] Cimino JJ, Bright TJ, Li J. Medication reconciliation using natural language processing and controlled terminologies. Medinfo 2007;12(Pt 1):679-83.
[19] Peters L, Bodenreider O. Using the RxNorm web services API for quality assurance purposes. AMIA Annu Symp Proc 2008:(accepted).

Legends

Figure 1. Graph of categories in RxNorm

Figure 2. Original representation of multi-ingredient drugs in RxNorm

Figure 3. Normalized representation of multi-ingredient drugs in RxNorm

Figure 4. Contrasting two paths (top: SCD=>SCDF=>SBDF=>SBD and bottom: SCD=>SBD=>SBDF). The path at the top violates one of the rules for meaningful paths and leads to irrelevant SDBF instances.

Figure 5. Exploring the path SCDC=>SCD=>SBD=>SBDC from the SCDC instance Warfarin 1MG

Table 1. RxNorm major categories

Table 2. List of equivalent paths in RxNorm with assessment of consistency among them

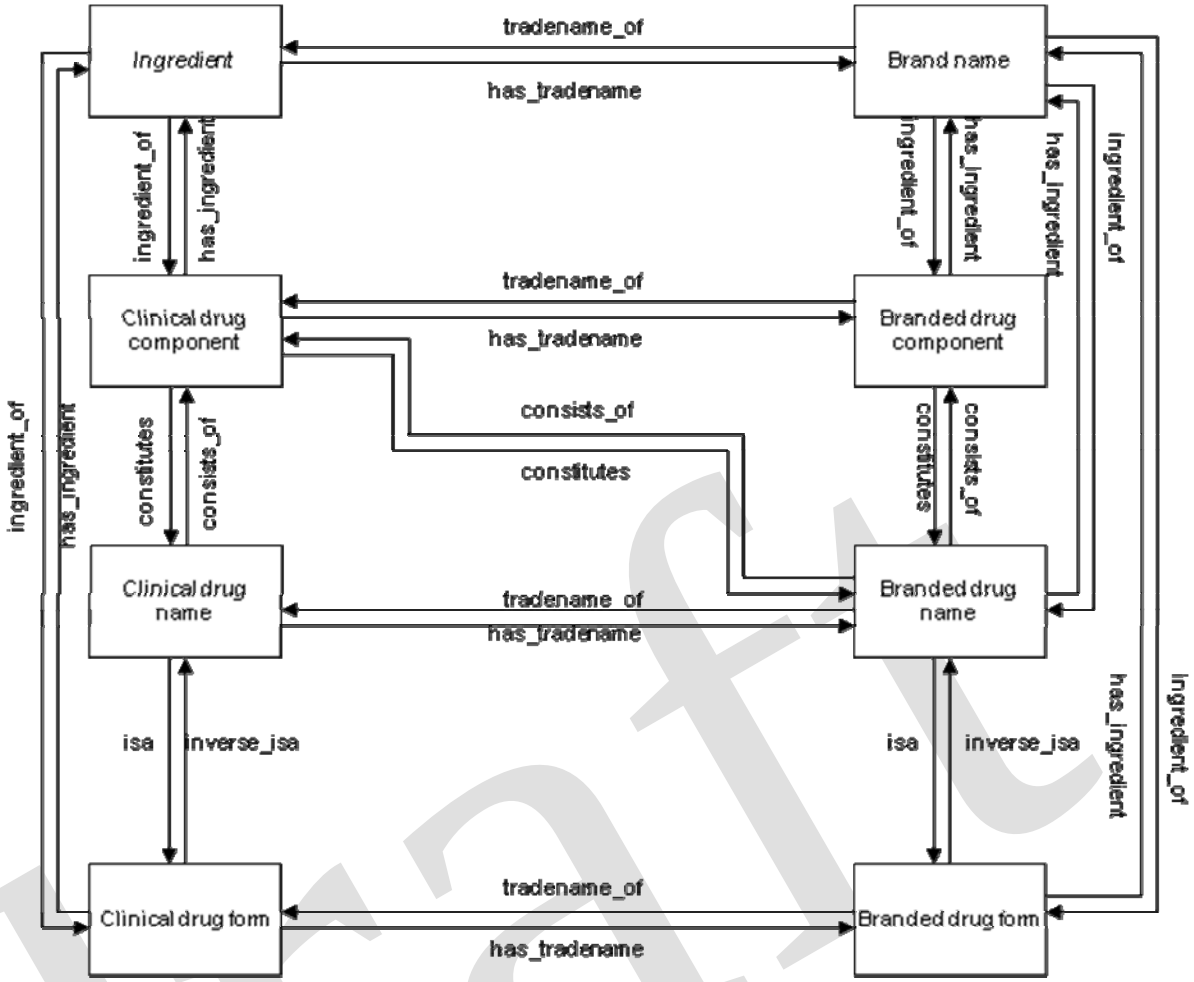

Figure 1. Graph of categories in RxNorm

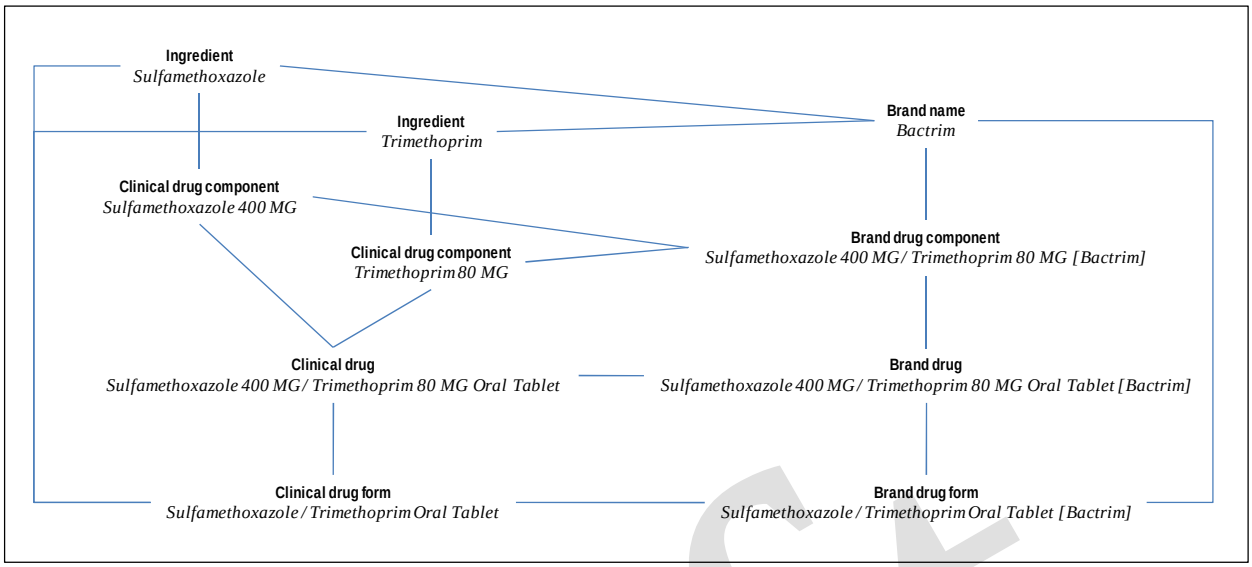

Figure 2. Original representation of multi-ingredient drugs in RxNorm

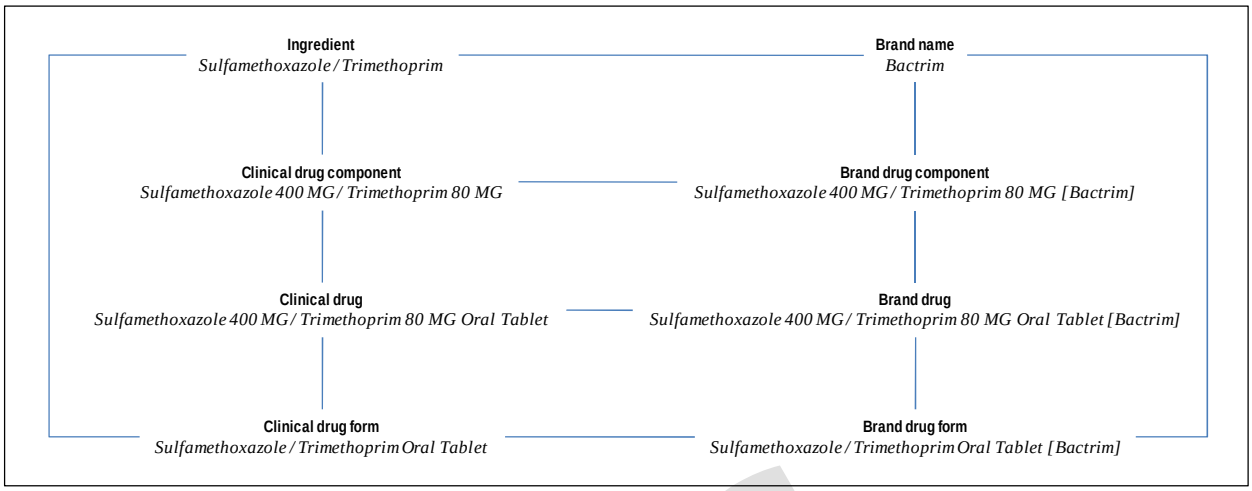

Figure 3. Normalized representation of multi-ingredient drugs in RxNorm

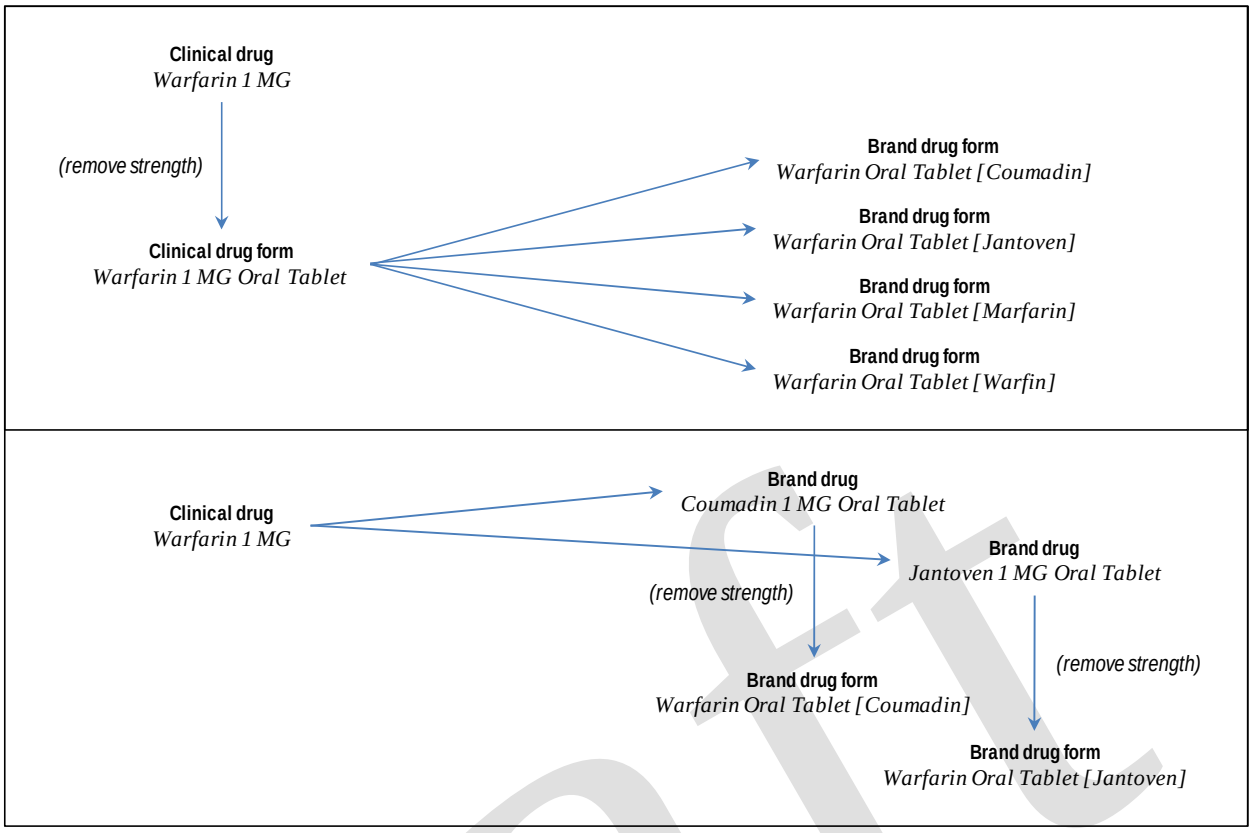

Figure 4. Contrasting two paths (top: SCD=>SCDF=>SBDF=>SBD and bottom: SCD=>SBD=>SBDF). The path at the top violates one of the rules for meaningful paths and leads to irrelevant SDBF instances.

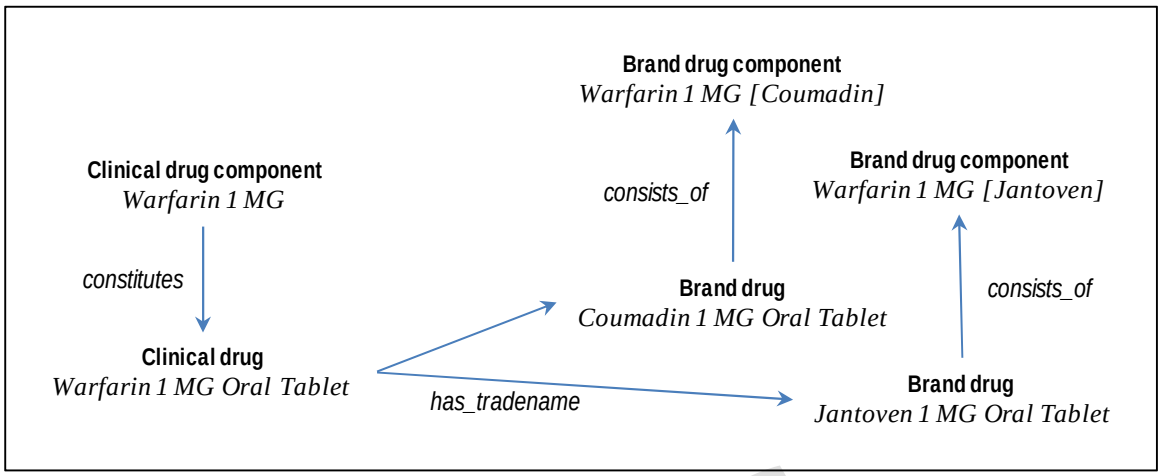

Figure 5. Exploring the path SCDC=>SCD=>SBD=>SBDC from the SCDC instance Warfarin 1MG

Table 1. RxNorm major categories

Category	Abbreviation	Instance
Ingredient	IN	Cetirizine
Brand name	BN	Zyrtec
Clinical drug component	SCDC	Cetirizine 5 MG
Branded drug component	SBDC	Cetirizine 5 MG [Zyrtec]
Clinical drug name	SCD	Cetirizine 5 MG Oral Tablet
Branded drug name	SBD	Zyrtec 5 MG Oral Tablet
Clinical drug form	SCDF	Cetirizine Oral Tablet
Branded drug form	SBDF	Cetirizine Oral Tablet [Zyrtec]

Table 2. List of equivalent paths in RxNorm with assessment of consistency among them

Start node	End node	Number of paths	Paths	Equiv. paths (not shown)	Target nodes reached	Error type
BN	SBDC	3	BN=>SBDC	2	13,868	
BN	SBD	3	BN=>SBD	2	14,539	
BN	SBDF	3	BN=>SBDF	0	11,376	
			BN=>SBD=>SBDF	1	11,340	1a
SBDC	SBD	1	SBDC=>SBD	0	14,539	
SBDC	SBDF	1	SBDC=>SBD=>SBDF	0	14,469	
SBD	SBDF	1	SBD=>SBDF	0	14,539	
IN	SCDC	2	IN=>SCDC	1	14,513	
IN	SCD	2	IN=>SCDC=>SCD	0	18,097	3
			IN=>SCDF=>SCD	0	18,097	
IN	SCDF	2	IN=>SCDF	0	8,160	
			IN=>SCDC=>SCD=>SCDF	0	8,104	1b,3
SCDC	SCD	1	SCDC=>SCD	0	18,097	
SCDC	SCDF	1	SCDC=>SCD=>SCDF	0	17,556	
SCD	SCDF	1	SCD=>SCDF	0	18,097	
IN	BN	25	IN=>BN	0	9,723	2a/b
			IN=>SCDC=>SBD=>BN	6	9,830	
			IN=>SCDC=>SBDC=>BN	7	9,830	
			IN=>SCDF=>SBDF=>BN	0	9,864	
			IN=>SCDC=>SCD=>SCDF=>SBDF=>BN	0	9,857	1b,3
			IN=>SCDC=>SCD=>SCDF=>SBDF=>SBD=>BN	1	9,831	1a,1b,3
			IN=>SCDF=>SBDF=>SBD=>BN	1	9,830	1a
			IN=>SCDF=>SCD=>SBD=>BN	2	9,830	3
IN	SBDC	11	IN=>BN=>SBDC	2	13,838	2a/b,2c
			IN=>SCDC=>SBDC	4	13,868	
			IN=>SCDC=>SCD=>SCDF=>SBDF=>SBD =>SBDC	0	13,869	1b,3
			IN=>SCDF=>SBDF=>SBD=>SBDC	0	13,868	
			IN=>SCDF=>SCD=>SBD=>SBDC	0	13,868	3
IN	SBD	11	IN=>BN=>SBD	2	14,509	2a/b
			IN=>SCDC=>SBD	4	14,539	
			IN=>SCDC=>SCD=>SCDF=>SBDF=>SBD	0	14,540	1b,3
			IN=>SCDF=>SBDF=>SBD	0	14,539	
			IN=>SCDF=>SCD=>SBD	0	14,539	
IN	SBDF	11	IN=>BN=>SBDF	0	11,355	2a/b,2c
			IN=>BN=>SBD=>SBDF	1	11,319	1a,2a/b
			IN=>SCDC=>SBD=>SBDF	4	11,340	
			IN=>SCDC=>SCD=>SCDF=>SBDF	0	11,369	1b,3
			IN=>SCDF=>SBDF	0	11,376	
			IN=>SCDF=>SCD=>SBD=>SBDF	0	11,340	1a
SCDC	BN	9	SCDC=>SBDC=>BN	4	13,868	
			SCDC=>SBD=>BN	3	13,868	
SCDC	SBDC	3	SCDC=>SBDC	2	13,868	
SCDC	SBD	3	SCDC=>SBD	2	14,539	
SCDC	SBDF	3	SCDC=>SBD=>SBDF	2	14,469	
SCD	BN	3	SCD=>SBD=>BN	2	14,539	
SCD	SBDC	1	SCD=>SBD=>SBDC	0	14,539	
SCD	SBD	1	SCD=>SBD	0	14,539	
SCD	SBDF	1	SCD=>SBD=>SBDF	0	14,539	
SCDF	BN	6	SCDF=>SBDF=>BN	0	11,376	
			SCDF=>SBDF=>SBD=>BN	1	11,340	1a
			SCDF=>SCD=>SBD=>BN	2	11,340	
SCDF	SBDC	2	SCDF=>SCD=>SBD=>SBDC	0	14,469	
			SCDF=>SBDF=>SBD=>SBDC	0	14,469	
SCDF	SBD	2	SCDF=>SBDF=>SBD	0	14,539	
			SCDF=>SCD=>SBD	0	14,539	
SCDF	SBDF	2	SCDF=>SBDF	0	11,376	
			SCDF=>SCD=>SBD=>SBDF	0	11,340	1a